Community analysis through semantic rules and role composition derivation

Community Analysis through Semantic Rules and Role Composition Derivation

Matthew Rowe[a,*], Miriam Fernandez[a], Sofia Angeletou[b], and Harith Alani[a]

[a] *Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA United Kingdom*
[b] *BBC Future Media & Technology, Dock House, Media City, Salford, M50 2LH United Kingdom*

Abstract

Online communities provide a useful environment for web users to communicate and interact with other users by sharing their thoughts, ideas and opinions, and for resolving problems and issues. Companies and organisations now host online communities in order to support their products and services. Given this investment such communities are required to remain healthy and flourish. The behaviour that users exhibit within online communities is associated with their actions and interactions with other community users while the role that a user assumes is the label associated with a given type of behaviour. The domination of one type of behaviour within an online community can impact upon its health, for example, it might be the case within a question-answering community that there is a large portion of expert users and very few users asking questions, thereby reducing the involvement of and the need for experts. Understanding how the role composition - i.e. the distribution of users assuming different roles - of a community affects its health informs community managers with the early indicators of possible reductions or increases in community activity and how the community is expected to change. In this paper we present an approach to analyse communities based on their role compositions. We present a behaviour ontology that captures user behaviour within a given context (i.e. time period and community) and a semantic-rule based methodology to infer the role that a user has within a community based on his/her exhibited behaviour. We describe a method to tune roles for a given community-platform through the use of statistical clustering and discretisation of continuous feature values. We demonstrate the utility of our approach through role composition analyses of the SAP Community Network by: a) gauging the differences between communities, b) predicting community activity increase/decrease, and c) performing regression analysis of the post count within each community. Our findings indicate that communities on the SAP Community Network differ in terms of their average role percentages and experts, while being similar to one another in terms of the dominant role in each community - being a *novice* user. The findings also indicate that an increase in *expert* users who ask questions and initiate discussions was associated with increased community activity and that for 23 of the 25 communities analysed we were able to accurately detect a decrease in community activity using the community's role composition.

Keywords: social web, communities, semantic web, behaviour, role analysis

1. Introduction

Online communities are now an integral part of the World Wide Web, they provide web users with the necessary environment in which they can interact and discuss topics of interest and seek answers to questions and support-requests. Such is the utility of online communities that companies now host discussion and support forums in order to support their products and services. Such usage reduces the need for consumers to contact telephone help desks as the necessary support information is instead provided by the community's users. A prime example of this is the UK telecommunications company BT who now provide a dedicated support community[1] in which customers can post their queries to new, emerging problems and find solutions to existing ones.

The investment in online communities, in terms of time, effort and money, means that community hosts and managers have a vested interest in the success of their community. This presents a clear need for the invested communities to remain healthy and active, thereby reducing the likelihood of the community's activity volume decreasing - i.e. a reduction in the number of posts - and maintaining the community's 'health'.

At present there is a limited understanding of how communities function and what leads to an increase or decrease in the health of a community. Communities are comprised of a mix of different users, many of whom exhibit differing behaviour and interact with one another in a disparate manner. One can regard communities as being online *ecosystems* where alterations in the behaviour of certain users can impact upon the community's dynamics, and therefore on its health. Earlier work by Preece [1] theorised that a community in which there was a single type of dominant behaviour would experience a decline in activity within the community and the subsequent churning of its users. This is comparable to a scenario in which a question-answering community, or one that is support driven, is largely comprised of expert users and with a low portion of users seek-

*Corresponding author. Tel: +44 (0)1908 655412
Fax: +44 (0)1908 653169
Email addresses: m.c.rowe@open.ac.uk (Matthew Rowe), m.fernandez@open.ac.uk (Miriam Fernandez), sofia.angeletou@bbc.co.uk (Sofia Angeletou), h.alani@open.ac.uk (and Harith Alani)

[1] http://community.bt.com/

ing answers. In this case one could imagine that expert users would reduce their activity as their utility diminishes with the lack of questions and problems being posed. We define user labels (e.g. expert, answer-seeker) as the *roles* that users assume within a given community. Users who have a role in one location, or in one community, may have a different role in another location. It may also be the case that as users develop and interact with a given community that over time their role changes, for example by going from a *newbie* to an *expert*.

The range of communities now being hosted and managed on the Web, at both the inter- and intra-platform levels, means that what may affect the health of one community may differ from another. Analysing the role composition of one community would provide an indication as to what worked for the community and what did not, allowing community managers to identify the role composition - i.e. the percentage breakdown of users assuming different roles - that functioned best or worst for their community.

Motivated by this setting we explore the following three research questions in this paper:

- *How does a change in its role composition affect the community?*
- *Are there different role compositions in differing communities? And what roles are dominant in disparate communities?*
- *Do distinct communities exhibit disparate patterns in how role compositions affect community activity?*

1.1. Contributions

In order to explore these research questions we devised an approach that facilitates the analysis of communities based on their role compositions, the approach is comprised of 3 stages: a) *modelling*, b) *role identification*, and c) *analysis*. The first stage *modelling* is where user behaviour is represented within a given context (i.e. community and time) using a behaviour ontology. The second stage *role identification* involves the derivation of roles for the community. For this stage we present a statistical-clustering based method that segments community users into distinct clusters and then aligns each cluster with a role. From this alignment semantic rules are then constructed that allow the roles of community users to be inferred based on their exhibited behaviour. The final stage *analysis* uses the semantic behaviour representation of a community's users together with the semantic rules to derive a community's role composition over time. This allows analyses to be made as to how the role composition correlates with community activity and how communities differ. Our contributions in this piece of work are four-fold:

- A behaviour ontology capable of representing user behaviour and context.
- A semantic rule-based approach to infer community roles.
- A method to align clusters to roles in a given platform using statistical clustering and discretisation of continuous feature values.
- Analysis of community role compositions on the SAP Community Network to: a) identify community differences; b) detect activity changes, and; c) predict post counts.

We have structured this paper as follows: section 2 describes the related work within the domains of role composition analysis and behaviour modelling. Section 3 describes the dataset that we used for our experiments and analysis from the SAP Community Network, a support-oriented community platform where users solicit help from community members on SAP-related product issues and technical problems. Section 4 describes the modelling and approach aspects of our work by detailing the representation of behaviour and roles and how we infer the role composition of a community using semantic rules. Section 5 presents the role identification stage of our approach in which we describe the use of statistical clustering to partition community users and align the clusters with role labels, thereby generating a set of roles for a given platform and the rules used to infer them. Section 6 describes the analysis of the SAP Community Network using role composition derivation and the experiments conducted. Section 7 presents the discussions of our findings and plans for future work and section 8 finishes the paper with our conclusions.

2. Related Work

2.1. Roles, behaviours and behavioural features

In this section we report on existing works that investigate behaviour patterns and role compositions in online communities. When investigating these topics it is key to have a clear understanding of what roles are, how they relate to human behaviours, and how these behaviours can be captured in terms of online community features.

A discussion about the definition of a role can be found in work by Golder and Donath [2]. In their discussion the authors state that a role can arise from the social context of a person and the dynamics of his/her relationships (e.g. the father family role) or from repeated interactions and agreements across practices (e.g. group planner, or decision-maker roles). In this work we focus on the second definition of role, identified as a set of behavioural patterns present in the social context of online communities. Examples of roles repeatedly mentioned in the literature are: *newbies*, *experts* or *lurkers*,

Each of these roles is identified by a set of behaviours, (or behaviour dimensions), such as engagement, contribution, popularity, participation, etc. The general procedure to model these behaviours in online communities is by translating them into measurable behavioural features from the social network graph with an associated intensity level (e.g. low, medium, high). For example, in the work of Hautz et al. [3] on the *Swarovski Enlightened Design Competition* online community,[2] three behaviour dimensions were identified; *motivation*, *attention grabbing*, and *idea generation*. These dimensions were measured by

[2] http://www.enlightened-jewellery-design-competition.com/

considering different combinations and levels (high, medium, low) of the features *in-degree*, *out-degree* and *number of designs uploaded*.

Similarly, Nolker and Zhou [4] identified three different behaviour dimensions; *spreading knowledge*, *keeping conversation going*, and *producing high conversation volumes*. These behaviour dimensions were measured from the combinations and levels (high, medium, low) of several social network features. The features they focused on were *degree* (number of conversations where the user has participated), *betweenness* (pairs of members who converse indirectly through another member), *closeness* (average conversation distance with all the other members of the community), and *discussion ratio* (percentage of one-way and two-way conversations). More recent approaches such as [5, 6] also modelled and computed behaviour dimensions by exploiting measurable features from the social network such as: *in-degree* (number of calls received from others), *out-degree* (number of calls made to others), *in-length* (total duration of calls received from others), *out-length* (total duration of calls made to others), and more complex social network graph measures such as *InnerPageRank* and *OuterPageRank*.

Mapping behaviour dimensions to specific community features is not a trivial task and is naturally dependent on the features that are of relevance to the community in question. Data preparation and feature computations often face problems of missing or inconsistent information [4]. It is therefore up to the community analyst to identify the correct and most appropriate metrics and features that can be used to measure behaviour dimensions in a given community.

A wide number of studies from different research communities (sociolinguistics, social psychology, ethnography communication, etc.) have aimed to capture the set of roles and behaviours present in online communities. For instance, Strijbos and Last [7] and Jenny Preece [1] defined the labels *captains* and *pillars*, *moderators* and *mediators* for those users who contribute with high intensity, reciprocity and persistence, and positive polarity to a community. Golder and Donath [2] labelled users who set the standard in a community as *celebrities*. Similar to the celebrity role are the roles: *Popular initiator*, *popular participant* and *joining conversationalist* [8] as their intensity, persistence and reciprocity are also quite high. Another type of prolific, but not as widely popular, user is the *elitist*, who demonstrates high values for the above dimensions but communicates with a smaller group of users.

Fisher and Smith [9] published one of the first works that provided '*operational definitions* - i.e. roles - based on their behavioural patterns. Looking into Usenet newsgroups the authors identified the *answer person*, who is engaged in many threads but usually posts once per thread and provides solutions and answers to other users enquiries and the *discussion person*, who, compared to the answer person, posts in less threads but has a higher persistence per thread, these contributions are considered to be conversational rather than responses.

The roles mentioned so far are associated with high community activity and, in general, positive community responses. Converse to this are those roles that are at the lower end of the activity scale. For instance the role *lurker* is the most frequently observed role in online communities and is defined as a participant who consumes but does not contribute and usually has a strong personal focus [2, 1, 7]. Similarly described roles are those of *content consumers* [10], *grunts* and *taciturns* [8] who do contribute but with low intensity. The polarity of the user contribution has also been used to distinguish the negative roles of *troll* and *flamer* who exhibit disruptive behaviour similar to the *ranter*. Like celebrities, ranters also demonstrate high intensity and persistence yet their primary goal is to raise discussions on the topic of their interest for some personal goals, same as *over-riders* and *generators* [7].

Despite the existing wide range of studies, there is still not a standardised or broadly accepted subset of roles and associated behaviours across communities. Indeed, some works like [4, 6] state that "*different communities have different needs and the roles that support these needs are therefore different*". However, although there is not a commonly agreed set of roles there is a tendency in the literature to reiterate certain behaviours like: popularity, engagement, contribution, initiation and focus. Based on these findings, our analysis aims to apply these *generic* set of identified behaviours to the SAP community and, without previous pre-conceptions, study which roles emerge from those dimensions.

2.2. Role Composition Approaches

Several role composition, or role identification, approaches have been reported in the literature. According to [11] these works can be divided in two general methodological approaches: interpretive analysis and structural analysis. Interpretive analysis approaches, such as the one proposed by Golder and Donath [2] employ methods like ethnography, content analysis, and surveys to capture behaviours and relations within groups. This is a prominent method used by anthropologists and sociologists in understanding groups and social systems. While highly useful in identifying and understanding important social roles and the context in which these roles develop, interpretive studies are very difficult to reproduce. This results in role definitions and findings that are difficult to compare across communities.

Structural analysis approaches [4, 5, 3, 8, 6] use formal methods like clustering or network structure analysis to identify relevant roles within the community. These approaches differ in their initial assumptions and in the methodology selected for the analysis. The work of [4] for example assumes the existence of: a) roles identified from the literature (leaders and motivators) and; b) a set of behavioural features identified from the social network graph including a set of well-known graph measures - e.g. *betweenness*, *closeness* - and their own adaptation of the TF*IDF measure.[3] They associate these behavioural features and their intensity level (high, moderate, low) to the preselected roles. When analysing the community they extract the roles based on the previous generated association.

[3] `http://en.wikipedia.org/wiki/Tf-idf`

The works of [3, 5] assume only the existence of a set of behavioural features extracted from the social network graph. In [3] the previously assumed behavioural features are *in-degree*, *out-degree* and the *number of designs uploaded*. Based on these features and their intensity level (high, medium, low) eight different roles are identified including: *motivator*, *attention attractor*, *idea generator*, *passive user*, etc. In the work of [5] the predefined behavioural features are two measures proposed by the authors from the network structure: the *InnerPageRank* and the *OuterPageRank*. Combinations of high and low values of these two features are used to represent the different roles. The main drawback of these approaches is that: either they use a very reduced set of behavioural features and represent each role with a simplistic combination of these features and their intensity level (high, medium, low) or, if the aim is to cover a broader set of features they need to limit the set of roles they aim to identify (otherwise the combinatorial options may increase significantly).

Overcoming this limitation [8, 6] assume a set of initial behavioural features and then perform cluster analysis to identify the set of roles that emerge from the community. Each cluster approximately corresponds to one role. While these approaches are based on formal cluster analysis, an informal observation of the clusters is performed afterwards in order to manually identify the roles and their associated behavioural features and values. In this paper we describe an approach that aims to support this *role identification* step with empirical data, such that the role labels attributed to a given cluster are derived from each cluster's behaviour dimensions and their distributions. We employ a maximum-entropy decision tree to generate the role labels without the need for a pre-conceived role collection. Another key difference of our approach with the aforementioned works is that while such works focus on identifying the key contributors of the community we aim to investigate a community's complete role composition without making any presumption of which users the administrators should pay more attention to. Under certain circumstances, like churn risk for instance, it would be better for administrators to identify not the key contributors but the users who are likely to leave the community.

2.3. Semantic Web and Role composition

According to Breslin et al. [12] "*At present online communities are islands that are not interlinked...*". Although this statement is from 2005 it still remains pertinent for today's Web given the myriad Social Web systems that support community development. Within this setting there are obvious commonalities across communities. For example, the same user may participate in several communities and even post the same content in those communities (e.g. people who link their Twitter and Facebook accounts so that the same status update is published on each). Establishing a semantic model allows better information sharing and interlinking, and would enable: a) analysis across communities; and b) better content search and recommendation - i.e. recommendation of items based on preferences that the user publicly defined in another network.

The work of Ankolekar et al. [13] is an example of the potential that a semantic model can bring to online communities by identifying and interlinking discussions and actions over the same objects, in this particular case software components. More recently, in 2010, Facebook announced the Open Graph protocol, which exploits RDF[4] to model and interlink users and objects within the Facebook social network. While these approaches have attempted to model and interlink objects and users within the same community, very few approaches in the literature have addressed the problem of representing the behaviour of users and their roles within online communities in a machine readable and shareable format. The most well-known ontology that addresses the problem of role definition is SIOC [12]. SIOC is written in RDF and is composed of eleven main classes: Community, Container, Forum, Item, Post, Role, Site, Space, Thread, UserAccount and Usergroup. The class `sioc:UserAccount` represents online community users and it is linked to the class `sioc:Role`, which represents the role that users may have within the Community. This ontology is based on, and reuses classes and relations from, several well-known ontologies such as the Friend Of A Friend (FOAF) vocabulary [14] and the Dublin Core Metadata Terms (dcterms).[5] In a more specific domain, software development, Ankolekar et al. [13] modelled a community ontology[6] to describe user roles: bug fixer, bug reporter, contributor, developer,etc.

Earlier work by Peter Mika and Aldo Gangemi defined an ontology for the representation of *social relations*,[7] enabling the strength of social ties to be defined and supporting social network analysis. This work was later on refined [15] to extend the traditional bipartite model of ontologies with the social dimension, leading to a tripartite model of actors, concepts and instances.

Additional and complementary work includes the study on modelling '*social reality*' performed by Hoekstra [16]. This work is motivated by experiences in the development of the LKIF Core ontology of basic legal concepts[8] and aims to model concepts for describing social reality: roles, beliefs, desires, obligations, permissions, intentions, etc. To model this social reality the context (time and place) is included in the design pattern, thereby representing the different role that a person may have depending on the context.

In this paper we extend existing work, in particular SIOC and the Social Reality ontology [16], to provide a dedicated behaviour ontology that models a given user and the behaviour that he/she exhibits within a given context - i.e. both time and community. In doing so we can support *role inference* such that a user's role in a given context can be derived through semantic rules. As we will demonstrate, this semantic rule-based approach allows the role composition of a community to be derived over time, thereby detecting the change in community composition and supporting community health analysis.

[4] `http://www.w3.org/RDF/`
[5] `http://dublincore.org/documents/dcmi-terms/`
[6] `http://www.cs.cmu.edu/~anupriya/community.owl`
[7] `http://www.cs.vu.nl/~pmika/research/foaf-ws/foaf-x.html`
[8] `http://www.estrellaproject.org/lkif-core/`

3. Dataset: SAP Community Network

To ground our work we use the SAP Community Network (SCN) for role identification and role composition analysis. The SAP Community Network is a collection of online forums hosted by SAP in which users can discuss SAP-related issues including software development, SAP products and usage of SAP tools. SCN contains a points-based reward system to encourage problem resolution within the community. Users post their problems or issues and SCN users then reply with possible answers or useful information for the original question poster. Points are then awarded by the question poster to the answer that he/she deemed to be the best one. Over time users therefore build up a reputation on the platform as being knowledgeable about certain subjects and topics by their ability to provide highly rated answers.

We were provided with a subset of the SCN covering 33 communities, listed in Table 1. The topics of the communities range from being concerned with a particular programming language - e.g. ABAP[9] General, where ABAP stands for Advanced Business Application Programming - through to support for SAP products - e.g. SAP Business One Core. The dataset contained 95,200 threads, 421,098 messages of which 78,690 were allocated points, and 32,942 users. As the post counts within Table 1 indicate, there is a large variance in activity between the communities. For instance community 264 (SAP Business One Core) has the highest activity with 85,057 posts and community 486 (Enterprise Social Systems) has the lowest with only 7 posts.

Figure 1 presents the daily post counts throughout the entire SAP dataset from the creation of the platform in 2004 through to early-2011 (when we were provided with the dataset). The plot shows a steady increase in activity over time with some marked degradation in activity in the latter quarter of 2010. This increase and decrease provides a useful test bed for our role composition experiments, given that we want to identify compositions that correlate with community activity both in terms of increases and decreases.

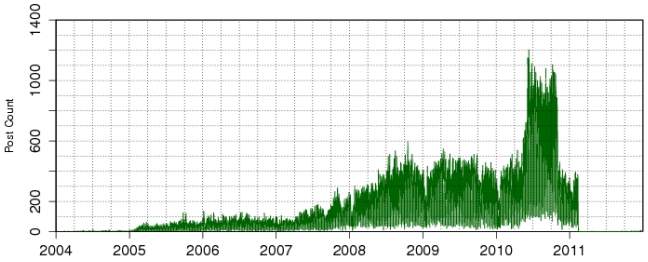

Figure 1: Number of Posts per day throughout the entire SCN dataset

[9]Advanced Business Application Programming (ABAP) is a programming language developed by SAP for their SAP Application Server.

Table 1: Communities and their IDs within the SCN dataset

ID	Name	Posts	Threads
101	Service-Oriented Architecture	9597	2570
161	SAP Business One Integration Technology	3163	812
197	Business Process Expert General Discussion	7464	2609
198	Business Process Modeling Methodologies	950	305
200	Organizational Change Management	230	47
201	Standards	367	163
210	Analytics	488	170
226	SAP Discovery System for Enterprise SOA	1105	408
252	SAP Business One E-Commerce and Web CRM	4487	1389
256	Governance, Risk and Compliance	19092	4279
264	SAP Business One Core	85057	17838
265	SAP Business One Product Development	2624	1127
270	Financial Performance Management General	8904	2482
281	Sustainability	190	42
319	Best Practice and Benchmarking	483	214
353	SAP Business One Reporting & Printing	38854	7744
354	SAP Business One Partner Solutions (Add-ons)	665	184
400	International Financial Reporting Standard (IFRS)	291	78
411	Operational Performance Management General	399	89
412	Busi' Planning & Consolidations: SAP NetWeaver	14439	3462
413	Busi' Planning & Consolidations: Microsoft Platform	18859	4245
414	SAP Strategy Management	1954	399
418	SAP Business One - SAP Add-ons	19656	3989
419	SAP Business One System Administration	16813	3222
420	SAP Business One Training	481	119
44	Process Integration	27768	4907
468	Green IT	39	8
470	Manufacturing Execution (ME)	1442	301
482	ASAP Methodology and Project Management	118	36
485	GS1 Standards and SAP	44	14
486	Enterprise Social Systems	7	3
50	ABAP, General	54718	13262
56	SAP Business One SDK	79800	18503

4. Modelling: Behaviour and Community Roles

User behaviour can change depending on the community in which the user is interacting and the time period. As a consequence the role that a user assumes is dependent on the context and can be different in the same community at different points in time and different at the same point in time yet within different communities. In this section we describe the representation of behaviour and context using our behaviour ontology and how semantic rules are used to infer the role that a user assumes given their exhibited behaviour and context.

4.1. Behaviour Dimensions

According to related work described in Section 2 the behaviour that users exhibit within differing types of online communities (e.g. discussion forums, question-answering platforms) can be described, in general, using six dimensions. In order to ground each dimension from an abstract notion of behaviour to something that is tangible in our assessed dataset (the SAP community network) we aligned each dimension, in a similar vein to existing work [3], with a specific feature that could be measured on the platform:

1. *Focus Dispersion:* the forum entropy of a user, where a high value indicates that the user disperses his/her activity across many SAP forums, while a low value indicates that the user concentrates his/her activity in a few forums. Let F_{v_i} be all the forums that user v_i has posted in and $p(f_.|v_i)$ be the conditional probability of v_i posting in forum $f_.$ - we can derive this using the post distribution of the user

- therefore we define the Forum Entropy (H_F) of a given user as:

$$H_F(\upsilon_i) = -\sum_{j=1}^{|F_{\upsilon_i}|} p(f_j|\upsilon_i) \log p(f_j|\upsilon_i) \quad (1)$$

2. *Engagement:* the proportion of users that the user has replied to. A larger value indicates that the user has contacted many different community members. Let Υ be the total number of users and $\Upsilon_{out,i}$ be users that υ_i has replied to, then the engagement of a user is defined as $\Upsilon_{out,i}/\Upsilon$.
3. *Contribution:* the proportion of thread replies that were created by the user. This measures the extent to which the user contributes replies to threads. Let P_r be the total set of replies authored by all users and $P_{r,i}$ be the set of replies authored by υ_i, we define the contribution of υ_i as $P_{r,i}/P_r$.
4. *Initiation:* the proportion of threads that were started by the user. This gauges how much the user instigates discussions and asks questions. Let P_s be set of thread starters authored by all users and $P_{s,i}$ be the set of thread starters authored by υ_i, we define the initiation of υ_i as $P_{s,i}/P_s$.
5. *Content Quality:* the average points per post awarded to the user. This provides a measure of expertise of the user. Let P_{υ_i} be the set of posts authored by υ_i and points(p) to be a function that returns the points awarded to post p, we define the content quality of υ_i as:

$$\frac{\sum_{j=1}^{P_{\upsilon_i}} \text{points}(p_j)}{|P_{\upsilon_i}|} \quad (2)$$

6. *Popularity:* the proportion of users that have replied to the user. A larger value indicates that the user is popular within the platform. Let Υ be the total number of users and $\Upsilon_{in,i}$ be the users that have replied to υ_i, then the popularity of a user is defined as $\Upsilon_{in,i}/\Upsilon$.

4.2. Behaviour Ontology

Analysing disparate communities on Social Web Systems provides insights into how behaviour differs between communities and how changes in behaviour can affect the development of different communities. A symptomatic problem of Social Web Systems, however, is the bespoke format that information is provided in. As we mentioned previously existing work by the Semantically Interlinked Online Communities (SIOC) project[10] has attempted to rectify this by providing a common format for information across communities through the SIOC ontology, describing user accounts, posts, forums and platforms. SIOC is focussed on providing a common semantic model for representing information across communities, and therefore it does not capture all the information required for measuring user and community behaviour.

User behaviour is contextual, how one person behaves in one context may differ from another, in essence they may behave differently in different locations or at different times. In our work we need to capture this contextual notion of user behaviour using the above dimensions, and then use this information to *identify* the user's role in a given community at a given point in time. For this purpose we have created the Open University Behaviour Ontology (OUBO),[11] a portion of which is shown in Figure 2. We regard this ontology as a natural extension to SIOC [12] that allows user behaviour to be captured over time and facilitate role inference.

Using an ontology and semantics to tackle the problem of behaviour analysis offers a number of advantages. Firstly, the ontology provides a generic, reusable, and machine understandable model for representing the concepts and properties required for describing user activities and measuring their behaviour. Secondly, due to the use of SIOC, this ontology greatly facilitates the integration of data from multiple social networking systems and data resources. Therefore the ontology can be used to measure behaviour of users across several community platforms. Thirdly, and most importantly, the ontology allows the rules for calculating behaviour (Section 4.3) to be seamlessly integrated with user data and behavioural labels and concepts. These advantages render the use of semantics to be a very practical and efficient approach.

4.2.1. Representing Behaviour

The primary information that we need to capture is a given user's behaviour, represented using the above numeric attributes for the behaviour dimensions. To do this we extend the SIOC ontology by providing a class called `oubo:UserImpact` in which we store the numeric behaviour attributes, this class is associated to the SIOC class `sioc:UserAccount` using the `oubo:hasUserImpact` predicate.

The class `oubo:UserImpact` models the impact of the user in a certain time period by storing, for that specific time frame, the value of the previously described behaviour dimensions. We also capture impact information related to posts using the `oubo:PostImpact` class such as the number of replies, comments and forwards that a post has had. Although we do not use this information within this paper, we have used it in other work that predicts the impact that a post will have on a community [17]. For this purpose we have also created the class `oubo:Post`, as as subclass of `sioc:Post` in order to capture post statistics.

4.2.2. Representing Roles

As we alluded to within the related work section there are various roles that are unique to specific types of Social Web Systems and certain roles that are common across such systems. We need to allow for bespoke role definitions depending on the platform and community under analysis. To enable such definitions we defined the class `oubo:Role` as a subclass of `sioc:Role` and `social-reality:OR`. This latter class is from work by Hoekstra [16] on abstracting roles in social contexts. The class `social-reality:OR` refers to an *Observer Relative*

[10] `http://sioc-project.org/`

[11] `http://purl.org/net/oubo/0.3` - we use the prefix `OUBO` hereafter for this namespace

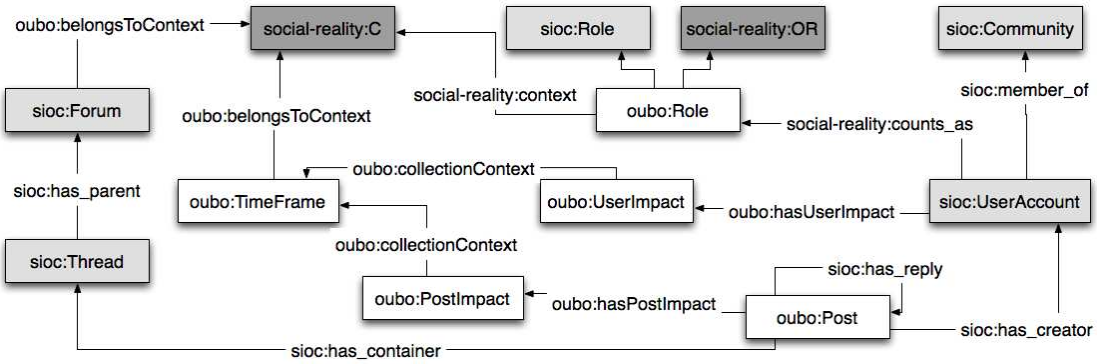

Figure 2: An overview of concepts and object properties from the Open University Behaviour Ontology (OUBO)

Fact (OR) which defines a *subjective* assessment, this could be a judgement or opinion formed by the assessor.

Analysts can then extend our ontology by defining specific specialisations of `oubo:Role` for the roles that they wish to infer. For instance in previous work by Chan et al [8] the authors used roles specific to a discussion message board (e.g. popular participant, grunt, etc.). Therefore by using our ontology the above class could be specialised for each role type and then individual users could be associated with the role they assume. The SAP semantic rules obtained as a result of this work are placed under this class and can be published online, allowing third parties to apply these rules for the derivation of role compositions for similar platforms.

4.2.3. Representing Context

There are two types of context that we wish to define: *location* and *time*. For the former we can use the SIOC classes `sioc:Forum` and `sioc:Community` to represent the community in which the user, defined as an instance of `sioc:UserAccount`, has been involved in. The class `sioc:Community` is a high-level concept that defines an online community. A community may consist of different types of objects (people, forums, sites, etc.). The class `sioc:Forum` represents a channel or discussion area in which posts are made. The class `sioc:Community` allows forums and sites to be grouped together under the same umbrella. For the latter context type (time) we created a class named `oubo:TimeFrame` that defines a given time period in which the user's behaviour statistics have been collected. We combine the temporal and location context aspects into a single context instance using the class `social-reality:C`, linking each respective class using the `oubo:belongsToContext` predicate. The class `social-reality:C` is also from Hoekstra's work on role abstraction and is used to represent a higher-level notion of context that can be used to include additional context information - i.e. time and locality in our case.

The above representations of behaviour, roles and context allow our approach to infer the role (`oubo:Role`) that a user (`sioc:UserAccount`) has in a given context (`social-reality:C`). We associate the user to their role using the predicate `social-reality:counts_as` and associate a given role to the context in which it applies by `social-reality:context`. Over time the role that a user assumes may change depending on the community in which they are interacting and time period. By providing abstractions of these aspects of context we can enable such inferences to be made, and capture the multiple roles that users may have at the same point in time but within differing locations and at the same location but within differing points in time. Statistical approaches, such as [4, 5, 3, 8, 6], do not allow for such adaptation and flexibility, and instead function over a specific dataset built from a specific time period.

4.3. Constructing Semantic Rules

Our approach to derive the role composition functions by taking the users who participated in the community over a given period of time and inferring the role of each user in the community, thereby providing a measure of the role composition - e.g. 10% roleA, 20% roleB, etc. We can then derive the role composition repeatedly over incremental time periods and capture how the composition changes in the community - in Section 6 we present how this information can be used to predict changes in a community's activity.

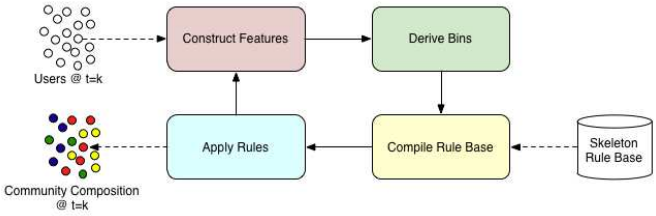

Figure 3: Overview of the approach to analyse user behaviour, label users with roles and derive the community composition

Figure 3 presents an overview of our approach for deriving a community's role composition over time. We begin by taking all the users within a community over a given time segment and calculating the features that describe the behaviour of each community user. Next we take the features used to measure the dimensions of behaviour and derive bins for each feature using *equal frequency binning*, this divides the range that a feature's value may take between three levels: *low*, *mid* and *high*.

This binning procedure performs *discretisation* and enables our approach to account for fluctuations in feature ranges between time steps. For instance, if we were not to use equal frequency binning and instead split a feature's range into thirds then we may produce a densely populated bin - e.g. low - that contains the majority of the population. Instead we wish to capture the notion of *relevance* where a *low* level for a feature is dependent on the community's population. By using *equal frequency binning* we take into account the underlying frequency distribution of the feature such that population density influences the boundary points for the feature levels.

The third stage of our approach then compiles the rule base from the *Skeleton Rule Base*. The *Skeleton Rule Base* contains a single rule for each role that is to be detected in the community. The antecedent of each rule contains a mapping between a feature and the level that that feature should be:

```
popularity=low, initiation=high -> roleA
```

The *Skeleton Rule Base* is platform-dependent and is set according to the analysis that is to be performed - in the following section we describe the process of building the *Skeleton Rule Base* and how the feature-to-level mappings are initially derived. The rules are constructed from the *Skeleton Rule Base* and the bins derived for each feature such that level boundaries are set within the rule:

```
popularity<0.5, initiation>0.4 -> roleA
```

The final stage of the approach is to apply the rules to the community users and infer each user's role. Rules are encoded using SPIN[12] functions that are triggered within the `WHERE` clause of a SPARQL `CONSTRUCT` query - we explain how rules are applied in the following subsection. Once every community user has been labelled with a role we can then derive the community's composition by the percentage of users that each role covers. The process of deriving the composition of a community can be repeated over time to detect changes in how the community evolves. In the analysis section (section 6) we demonstrate how the role composition of a community can be used to detect behavioural differences between disparate forums and predict changes in a community's activity.

4.4. Applying Semantic Rules

As we mentioned above, our rules used to infer the role of individual community users are encoded using SPIN functions. Other alternatives were considered, such as SWRL[13] and RIF.[14] Several discussions can be found[15] about the characteristics, differences and advantages of each of these rule representation languages. The goal of RIF is to create an interchange format for use between rules engines. As such, unlike SPIN, RIF is not specifically or particularly aligned with RDF. More importantly, SPIN is based on SPARQL, which makes it a more expressive language than SWRL or RIF.[16] Another advantage of SPIN, with respect to the two rule languages mentioned above, is that rules are written in SPARQL, a familiar language for most RDF users, and a language that makes the rules portable, not across rules engines, but across RDF stores. SPIN rules can be directly executed on the data stores and no intermediate engines with communication overhead need to be introduced. Finally, full support for SPIN is provided by TopBraid, including the TopBraid composer editor, templates,etc.[17]

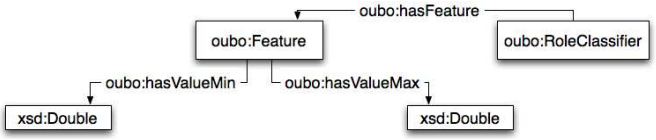

Figure 4: Association of Roles with Features

Below we show an example SPARQL `CONSTRUCT` query used to infer a user's role within a community. Within the rule's `WHERE` clause there are two SPIN functions: `oubo:fn_getRoleType(?user,?temp,?forum)` and `smf:buildURI("oubo:Role{?type}")`. The former function takes as parameters the user whose role is to be inferred - designated by `?user`, the time period over which the user's behaviour is to be assessed - given by `?temp` - and the location in which the user's behaviour to be assessed and role inferred - given by `?forum`.

```
PREFIX oubo:<http://purl.org/net/oubo/0.3> .
PREFIX sioc: <http://rdfs.org/sioc/ns> .
PREFIX smf: <http://topbraid.org/sparqlmotionfunctions#> .
PREFIX social-reality: <http://purl.org/net/social-reality#> .
CONSTRUCT {
    _:role a ?t .
    ?user social-reality:count_as _:role .
    _:context a social-reality:C .
    _:role social-reality:context _:context .
    ?forum oubo:belongsToContext _:context .
    ?temp oubo:belongsToContext _:context
} WHERE {
    BIND(oubo:fn_getRoleType(?user,?temp,?forum) AS ?type) .
    BIND(smf:buildURI("oubo:Role{?type}") AS ?t)
}
```

The behaviour information describing the user is then looked up - i.e. returning an instance of `oubo:UserImpact` for `?user` at a given time point and for a given location - and each of the rules are applied over the behaviour features until one matches. Each rule is defined as an instance of `oubo:RoleClassifier` and is associated with a set of features as shown in Fig. 4. Each feature has a minimum and maximum value which specify the range of feature values a user should have for this feature in order to be assigned to this role. We use the skeleton rule of the role to provide the rule's syntax and then replace the levels with the necessary bounds produced by our binning procedure.

The `?type` variable returns the role label for the user and the second function (`smf:buildURI("oubo:Role{?type}")`) constructs a Uniform Resource Identifier (URI) for the given role which is then bound to `?t` using the SPARQL 1.1 function `BIND`. This returns the URI of the role that should be assigned to the user (`?user`).

[12] `http://spinrdf.org/spin.html`

[13] `http://www.w3.org/Submission/SWRL/`

[14] `http://www.w3.org/TR/rif-core/`

[15] `http://spinrdf.org/faq.html,http://topquadrantblog.blogspot.co.uk/2011/06/comparing-spin-with-rif.html`

[16] `http://www.w3.org/2005/rules/wiki/RIF_FAQ`

[17] `http://www.topquadrant.com/products/SPIN.html`

Within the CONSTRUCT clause of the SPARQL query we then build the relation between the user and his/her role that has been inferred. The first line defines a blank node (_:role) as being an instance of the inferred role (?t). The user (?user) is then assigned to the role (?t) using social-reality:count_as. The _:context in which the role is applicable is defined as an instance of social-reality:C and is attributed to the _:role using the social-reality:context predicate. The location (?forum) and temporal (?temp) context information is then associated with ?context using the oubo:belongsToContext predicate.

Following this process we can perform SPARQL queries to retrieve all the roles that a given user has had, the cycle between roles that a user has exhibited over time and the composition that a given community has at a given point in time, along with how this changes.

5. Role Identification: Tuning Roles

Compilation of the *Skeleton Rule Base* is a platform-dependent process as distinct types of Social Web Systems contain certain community roles - e.g. discussion message boards contain conversation-driven roles, microblogging platforms contain celebrity users, etc. Decisions must be made as to what roles to monitor in a given community and whether those roles are appropriate. In this section we describe the compilation of a *Skeleton Rule Base* for roles that users assume on the SAP Community Network. We use a combination of statistical clustering and manual inspection to perform this *role identification* step by partitioning the community's users into clusters, deriving feature-to-level mappings for each cluster and then aligning clusters with role labels.

5.1. Tuning Segment Selection

In order to cluster users into distinct roles we needed to select a time segment from the SAP Community Network over which we could perform *tuning*, this section needed to be distinct from our later analysis experiment in order to ensure independence. To do this we recorded the number of posts that were published on the platform every day throughout the entire dataset (described in section 3) - this is shown in Figure 1. We then assessed the distribution of posts throughout the duration of the platform, seeking a 6 month portion of the data - using this duration based on prior work by [8] - over which we could perform clustering. When selecting this tuning segment we noted that prior to 2008 there is markedly less activity than post 2008 and that there is also a large spike in activity throughout the latter half of 2010. We also wanted to ensure that we had a sufficient period over which we could perform our later analysis, given that we could not include the *tuning* segment in this portion and that the *tuning* segment must appear before the *analysis* section. Therefore for the *tuning* segment we selected the first 6 months of 2008 and used the remaining data - i.e. post the second half of 2008 - as our *analysis* section.

5.2. Identifying Correlated Behaviour Dimensions

In order to identify distinct community roles via clustering we need to be able to interpret key differences between the clusters. The aforementioned behaviour dimensions, although intended to be distinct, may in fact be correlated. We need to detect these correlated dimensions so that they can be removed and the dimensionality of our dataset reduced, thereby aiding discrimination between roles. To do this we built the above behaviour dimensions, and therefore the assigned features, for each user in our tuning dataset and then measured the Pearson correlation coefficient (r) between each dimension. Table 2 shows the correlation coefficient between each dimension within the dataset. In order to filter out the highly correlated dimensions that were significant we ran the Pearson correlation coefficient significance test where $r > 0.75$. In Table 2 we have marked all correlations that are significant at $\alpha = 0.01$. We found that *engagement*, *contribution* and *popularity* were all highly correlated with one another. Therefore we removed the first two dimensions from our dataset, resulting in the following dimensions remaining: *focus dispersion*, *initiation*, *content quality* and *popularity*.

5.3. Clustering Roles

Following the filtration of the initial dimensions we are left with dimensions that are distinct from one another, this forms the basis for clustering users into roles. By dividing users into distinct groups we attempt to separate those users based on their behaviour and therefore discover distinct roles on the platform. Several clustering methods exist from the literature, we therefore need to select the method that achieves the best clustering, thereby performing *model selection*. We ran three different unsupervised clustering algorithms: Expectation-Maximization (EM) [18], K-means [19] and Hierarchical Clustering,[18] over the 6-months' tuning segment of data. Each of these approaches requires the number of clusters k to be provided as an *a priori* parameter. The *model selection* phase not only requires choosing the correct clustering method but also selecting the optimum number of clusters to use. To judge which model performs best - i.e. cluster method and number of clusters - we make this selection based on the *cohesion* and *separation* of a given clustering, in essence we want to optimise the following two criteria:

1. *Maximise the intra-cluster similarity*
2. *Maximise the inter-cluster dissimilarity*

For each clustering algorithm (Ψ) we iteratively increase the number of clusters (k) to use where $2 \geq k \geq 30$. At each increment of k we record the *silhouette coefficient* produced by Ψ, this is defined for a given element (i) in a given cluster as:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{3}$$

Where a_i denotes the average distance to all other items in the same cluster and b_i is given by calculating the average distance with all other items in each other distinct cluster and then

[18] http://en.wikipedia.org/wiki/Hierarchical_clustering

Table 2: Correlation Coefficients of dimensions with significant correlations marked for $r > 0.7$

	Dispersion	Engagement	Contribution	Initiation	Quality	Popularity
Dispersion	1.000	0.277	0.168	0.389	0.086	0.356
Engagement	0.277	1.000	0.939**	0.284	0.151	0.926**
Contribution	0.168	0.939**	1.000	0.274	0.086	0.909**
Initiation	0.389	0.284	0.274	1.000	-0.059	0.513
Quality	0.086	0.151	0.086	-0.059	1.000	0.065
Popularity	0.356	0.926**	0.909**	0.513	0.065	1.000

taking the minimum distance. The value of s_i ranges between -1 and 1 where the former indicates a poor clustering where distinct items are grouped together and the latter indicates perfect cluster cohesion and separation. To derive the silhouette coefficient ($s(\Psi(k))$ for the entire clustering we take the average silhouette coefficient of all items. The coefficient provides a measure of the *quality* of the clustering by considering the cohesion (i.e. how similar intra-cluster items are to one another) and the separation (i.e. how dissimilar inter-cluster items are) in a produced clustering.

Figure 5 shows the tuning of each clustering algorithm when the number of clusters is increased. We find that the best clustering model and number of clusters to use is K-means with 11 clusters. The plot indicates that for smaller cluster numbers ($k = [3, 8]$) each clustering algorithm achieves comparable performance, however as we begin to increase the cluster numbers K-means improves while the two remaining algorithms produce worse cohesion and separation. The reason for this is the method of iterative assignment and updating that K-means employs by inducing initial means, mapping the closest items - based on Euclidean distance in our case - to those means and then updating the means based on the item assignment. This process is repeated until no new assignments can be made. This method allows distinct roles to be captured that may not cover many users in the dataset, while EM and Hierarchical clustering produce groupings which ignore these distinct roles.

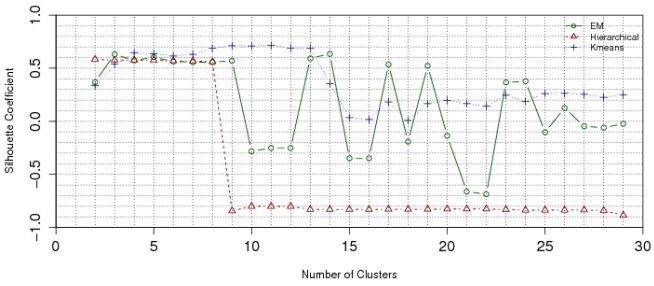

Figure 5: Clustering algorithms run with increasing cluster numbers, includes the silhouette coefficient measured at each epoch

5.4. Role Labelling

Existing work by Chan et al. [8] performed role labelling by clustering users from a discussion message board and then inspecting the clusters to see which role labels from the literature they resembled. We extend this work by providing an empirical basis for role labelling which makes no *a priori* assumptions of role labels and instead derives the labels according to the dimensions and levels in each cluster. Role label derivation first involves inspecting the dimension distribution in each cluster and aligning the distribution with a level mapping (i.e. *low*, *mid*, *high*). This enables the conversion of continuous dimension ranges into discrete values which our semantic rule-based approach requires in the *Skeleton Rule Base*. To perform this alignment we assess the distribution of each dimension and derive boundary points for the three feature levels using an equal-frequency binning approach.

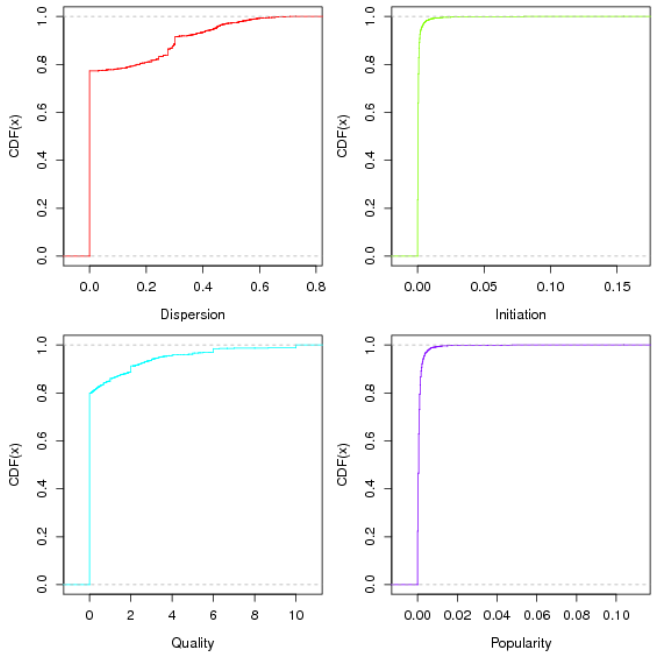

Figure 6: Cumulative density functions of each dimension showing the skew in the distributions for initiated and in-degree ratio

Figure 6 shows the empirical cumulative density functions for each dimension in our tuning sample. A large portion of the *dispersion* (i.e. entropy) distribution (78%) is found to be 0 indicating that these users always post in the same forum and do not deviate away, at the other extreme very few users are found to post in a large range of forums. For *initiation* and *popularity* the density functions are skewed towards low values where only a few users initiate discussions and are replied to by large portions of the community. *Quality* is also skewed

Table 3: Mapping of cluster dimensions to levels. The clusters are ordered from low patterns to high patterns to aid legibility.

Cluster	Dispersion	Initiation	Quality	Popularity
1	L	L	L	L
0	L	M	H	L
6	L	H	M	M
10	L	H	M	H
4	L	H	H	M
2,5	M	H	L	H
8,9	M	H	H	H
7	H	H	L	H
3	H	H	H	H

towards lower values indicating that the majority of users do not provide the best answers consistently. These plots indicate that feature levels derived from these distributions will be skewed towards lower values, for instance for *initiation* the definition of *high* for this feature is anything exceeding $1.55\text{x}10^{-5}$.

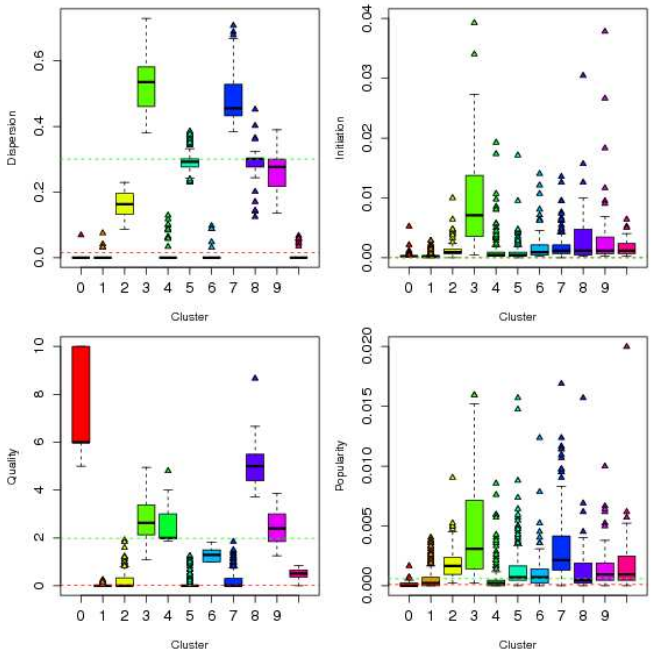

Figure 7: Boxplots of the feature distributions in each of the 11 clusters. Feature distributions are matched against the feature levels derived from equal-frequency binning

The distribution of each dimension is shown in Figure 7 for each of the 11 induced clusters. We assess the distribution of each feature for each cluster against the levels derived from the equal-frequency binning of each feature, thereby generating a feature-to-level mapping. This mapping is shown in Table 3 where certain clusters are combined together as they have the same feature-to-level mapping patterns - i.e. 2,5 and 8,9.

In order to derive the role labels for each cluster we use a maximum-entropy decision tree to divide the clusters into branches that maximise the dispersion of dimension levels. Figure 8 shows the separation of the clusters from a complete grouping into a single cluster, or merged clusters in the case of

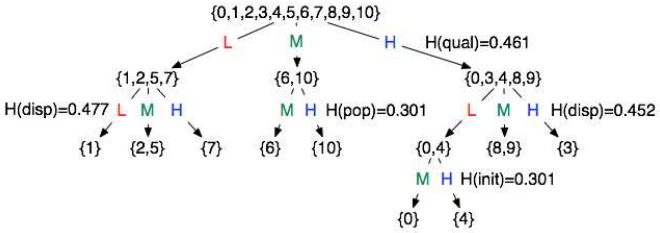

Figure 8: Maximum-entropy decision tree used to segment the clusters into minimal-distance paths. The paths are used to generate the role labels for each respective cluster.

5,7 and 8,9, in each leaf. To perform the separation at a given decision node, we measure the entropy of the dimensions and their levels across the clusters, we then choose the dimension with the largest entropy. This is defined formally as:

$$H(dim) = -\sum_{level}^{|levels|} p(level|dim) \log p(level|dim) \quad (4)$$

At the root node - i.e. the top node containing all clusters - we find the maximum-entropy dimension to be *quality* achieving an entropy of $H(x) = 0.461$. After separating the clusters into the respective branches for each dimension level we then assess the dispersion of dimension levels within each branch. Choosing the left split containing clusters 1,2,5,7 we find that *dispersion* yields the highest entropy where $H(x) = 0.477$ and divide the clusters up according to this dimension and their respective levels.

We perform this process until single clusters, or the previously merged clusters, are in each leaf node and then use the path to the root node to derive the label. For instance, for cluster 0 the path from the root node to the leaf node is *quality=high, dispersion=low, initiation=medium*, thereby deriving the role label **Focussed Expert Participant** for the cluster. In the label *focussed* describes the focus dispersion of the role - i.e. it is low and therefore not distributed, *expert* describes the level of expertise that a user will have - i.e. being high given the quality of their answers - and *participant* denotes the extent to which this role starts threads - i.e. being in the middle in this case and thus being both an initiator and an answerer.

By using entropy to assess which dimension to split the clusters we account for the largest variance in the clusters according to the dimension levels. This therefore derives the shortest role labels given that we generate the *purest* split possible at each branch and therefore reduce the depth to which the tree must be grown. Based on this method of deriving the role labels using dimension splits we produced the following role labels for each cluster from Table 3:

- **1 - Focussed Novice**: this user is focussed within a few select forums but does not provide good quality content.
- **0 - Focussed Expert Participant**: this user type provides high quality answers but only within select forums that they do not deviate from. They also have a mix of asking questions and answering them.

- **6 - Knowledgeable Member**: has medium-level expertise (i.e. he/she is neither an expert nor a novice) and has medium popularity
- **10 - Knowledgeable Sink**: user who has medium-level expertise but who gets a lot of the community replying to them - hence *a sink*. Differs from cluster 6 in terms of popularity.
- **4 - Focussed Expert Initiator**: similar to cluster 0 in that this type of user is focussed on certain topics and is an expert on those, but to a large extent starts discussions and threads, indicating that his/her shared content is useful to the community
- **2, 5 - Mixed Novice**: is a novice across a medium range of topics
- **8,9 - Mixed Expert**: medium-dispersed user who provides high-quality content
- **7 - Distributed Novice**: participates across a range of forums but is not knowledgeable on any topics
- **3 - Distributed Expert**: an expert on a variety of topics and participates across many different forums

The derived role labels above can be added to the ontology and consequently used to read and track the behaviour of individual users and communities over time.

6. Analysis: Community Health

Deriving a community's role composition provides community operators and hosts with a *macro-level* view of how their community is functioning. Understanding what is a *healthy* and *unhealthy* composition in a community involves analysing how a given role composition has been associated with community activity, interaction or some other measure in the past and reusing that knowledge. Forums and communities operating within the same platform may also differ such that what turns a community healthy in one location may be different from another. In this section we describe how community analysis is possible through our presented approach to derive the role composition of a community using semantic rules.

6.1. Experimental Setup

To demonstrate the utility of our approach we analysed 25 of the 33 SAP communities from 2009 through to 2011, removing 8 communities with <100 threads in the analysis window - previous experiments found these forums to be outliers. Figure 9 shows how our dataset was divided into the *tuning* section - i.e. the first half of 2008 in which we derived our clusters and aligned them to roles (as described in Section 5) - and the *analysis* section. We began with 1st January 2009 as our *collect date* by taking a *feature window* 6 months prior to this date (going back to the 2nd half of 2008) in which we measured the behaviour dimensions for each community's users. In order to gauge the role composition in a community over time we move our *collect date* on one week at a time and use the 6-months prior to this date as our *feature window*. As Figure 9 demonstrates we repeat this process until we reach 2011.

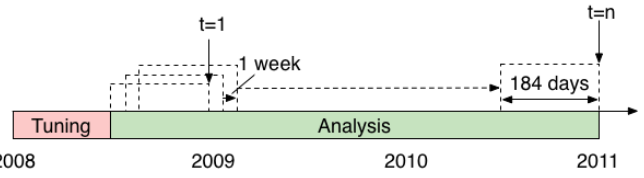

Figure 9: Windows used for a) tuning of the clusters and the derivation of roles and b) the analysis of community health. Role composition is derived every week from 2009 onwards using a 6-month window going back from the collection date.

By measuring the behaviour dimensions of individual users in individual communities we are able to infer the roles of the users using the semantic rules described in Section 4. This provides a *micro-level* assessment of the roles that individual users assume. We can then look at the *macro-level* by deriving the role composition of a given community at a given point in time by measuring how many users have a specific role. Such role composition analysis allows for predictions to then be made. To demonstrate the application of such analysis we performed three distinct experiments (each designed to explore one of our three aforementioned research questions):

1. *Composition Analysis:* assesses the average role composition in each community and clusters communities based on the compositions - allowing us to explore the research question: *Are there different role compositions in differing communities?* We also pick out each community's most popular role and measure what percentage of the community that role covers, thereby exploring the second half of the above research question: *And what roles are dominant in disparate communities?* We also assess the differences between communities based on the distribution of experts.
2. *Activity Increase/Decrease:* we perform a binary classification task to detect either an increase or decrease in community activity based on its role composition, exploring: *How does a change in its role composition affect the community?* We formulate this experiment such that at timestep $t = k + 1$ we predict whether the community's activity (i.e. number of posts) has *increased* or *decreased* since $t = k$. For features we use the 9 roles and the values are given by their percentages at $t = k + 1$. We train a logistic regression classifier and a J48 decision tree classifier and perform 10-fold cross-validation. We choose the best performing model according to F_1 values and plot the ROC curves to show the differences in performance between the communities.
3. *Post/User Count Regression:* we perform two linear regression analyses. The first analysis regresses the role composition of an individual community on the post count observed within the feature window. We measure the coefficient of determination (R^2) value to gauge the model fit

and report on the coefficients and how they differ between communities, thereby exploring our third research question: *Do distinct communities exhibit disparate patterns in how role compositions affect community activity?*
For the second analysis task we explore the relation between community size and role compositions. We induce a single regression model for all SCN communities by regressing the role composition on the user count and report on the model's fit, using the coefficient of determination. We then assess the correlation between increased community size and roles.

6.2. Results: Composition Analysis

We used the average role composition of each SAP community as its *composition motif* and these motifs as vectors to describe each SAP community. By running a Principal Component Analysis (PCA) over the data we grouped communities together that exhibited similar role compositions. Figure 10 shows the PCA plot and how disparate communities were grouped together.

We found **226 (Business Process Expert General Discussion)** and **319 (Best Practice and Benchmarking)** grouped together indicating that similar compositions occur in both these communities. Likewise, **161 (SAP Discovery System for Enterprise SOA)** and **265 (SAP Business One Product Development)** were clustered together. These two communities deal with discussions and topics related to SAP Business One (resource planning software) and how components can be developed and integrated into the platform. The PCA plot in Figure 10 shows that forum **470 (Manufacturing Execution)** is placed towards the top-centre of the graph. In this case the community exhibits a distinct composition from the other communities.

The principal component analysis shows what communities are similar to one another in terms of average compositions. It does not, however, indicate *how* the compositions differ. The latter part of our second research question asked *What roles are dominant in disparate communities?* To explore this question, and provide an insight into how the communities' compositions actually differ, we identified the most popular role in each community from its *composition motif* and measured the percentage that that role covered. This is shown in Figure 11 where the average role composition for each community is shown with the percentage breakdowns for each role.

Figure 11 indicates that the most dominant role across the communities is **Focussed Novice**. This role is assumed by those users who are non-experts and are enquiring about a specific topic. They are focussed in that they concentrate their posting behaviour in a few selected forums and do not deviate from those. **Distributed Novice** and **Mixed Novice** are then evenly distributed across the remaining forums. The former role is assumed by users who post in many different forums and are non-experts, while the latter role is for users who post in a medium number of forums (i.e. between the low and high thresholds for *focus dispersion*). For all the communities analysed the dominant role was found to be one involving non-expert users, thereby indicating, unsurprisingly, that SAP communities are comprised of users looking for an answer to a given problem.

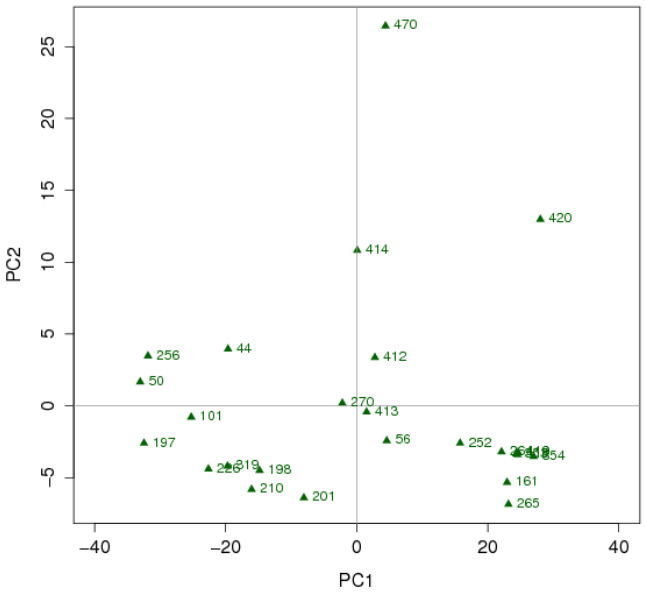

Figure 10: Principal Component Analysis (PCA) plot of each community by their average role composition

The utility of support communities is dependent on the experts within such forums that provide answers and help users solve their problems. When identifying the roles for the SAP Community Network we found four expert roles: **Focussed Expert Participant**, **Focussed Expert Initiator**, **Distributed Expert** and **Mixed Expert**. To delve deeper into the compositional differences between the SCN communities we plotted the percentage of users in each community that assumed these expert roles in Figure 12.

Focussing on the previously described communities that were either clustered together or isolated in the PCA plot (Figure 10) we find similar patterns in the expert roles. For instance, for **226 (Business Process Expert General Discussion)** and **319 (Best Practice and Benchmarking)** Figure 12 shows that both communities' have low proportions of **Focussed Experts** and instead have higher levels of **Mixed Experts** and **Distributed Experts**. The nature of the communities being associated with general discussions surrounding SAP technologies suggests that focussed experts are less likely to function in such environments.

For **161 (SAP Business One Integration Technology)** and **265 (SAP Business One Product Development)** like forums 197 and 50, we find low levels of **Focussed Experts**, in each case being almost 0. While for **Distributed Experts** the role percentages for 161 and 365 are relatively high. Each forum deals with problems concerning the SAP product '*Business One*' and, as we describe in the dataset section of this paper, there are also several other forums that relate to this product. The reason for the distributed experts in these two forums is due to those experts spreading their activity over the other fo-

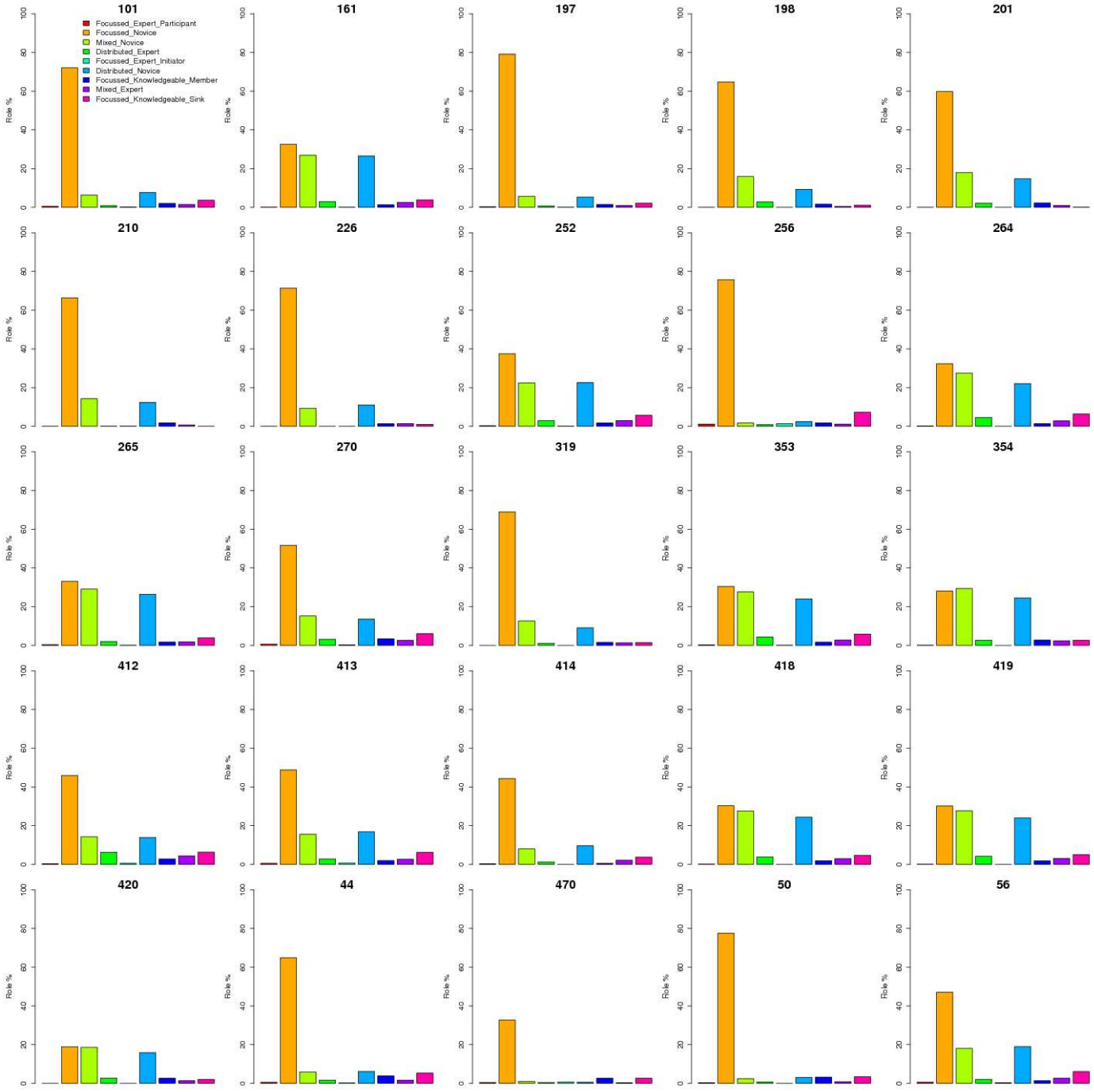

Figure 11: Bar charts of the roles in each community and the average percentage of users that that role covers

rums that concern SAP Business One, as a result they are not focussed on one distinct community.

The community that was placed towards the top-centre of the PCA plot, **470 (Manufacturing Execution)** deals with a relatively distinct topic, when considering the nature and subject of the other SAP communities. As a result we find that 470, according to Figure 12 has a high percentage **Focussed Experts** relative to the other communities, while having one of the lowest percentages of **Distributed Experts**. The latter of these findings is due to expert users who have knowledge in the area of manufacturing being less likely to participate in the other forums in our dataset due to their distinct topics. The distinct topical nature of forum 470 is confirmed by the high percentage of **Focussed Novice** users, which was the highest among all communities, found when inspecting the remaining roles from Figure 11.

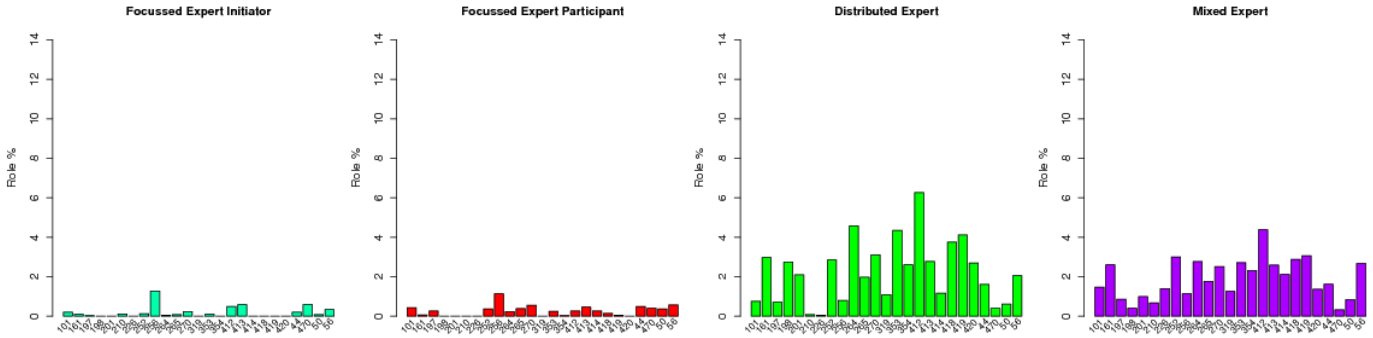

Figure 12: Bar chart of the dominant role in each community and the percentage of users that that role covers

6.3. Results: Activity Increase/Decrease Prediction

As community hosts and operators invest a lot of money, time and effort into maintaining online communities, a notable effect that they wish to avoid is a drop in activity. Activity can be regarded as a basic signifier of health such that if activity is reduced then the interaction and usage of the community has also diminished, while an increase could indicate that the community is becoming popular. Changes in behaviour in the community may affect community activity, and therefore its health and evolution. Based on this hypothesis we investigate the following question: *How does a change in its role composition affect the community*

To detect increases and decreases in community activity we tested the performance of the logistic regression and J48 decision tree classifiers by setting the class label based on either an *increase* or *decrease* in activity since the last time step and using the current time step's role composition as features. In doing so we could examine whether role compositions could be used to detect any changes in the community's health, measured by activity. Table 4 presents the results from this detection task when assessed using 10-fold cross validation. Out of the two classification models that we tested logistic regression achieved the best performance by outperforming the J48 decision tree in terms of recall and f-measure, while the Kappa statistic shows the achieved improvement over the random classifier. The higher recall level indicates that using this classification model allows community hosts to detect a larger portion of the activity change than using the J48 decision tree.

Table 4: Features used for our analysis including user features (first section), content features (second section) and focus features (third section)

Model	κ	Precision	Recall	F_1
Logistic	0.291	0.689	0.700	0.681
J48	0.263	0.676	0.687	0.677

The results indicate that we only yield satisfactory values for precision and recall when using either classifier, producing 0.689 and 0.700 for the best performing logistic regression classifier for each respective measure. One possible explanation is that frequent fluctuations in activity could impact upon either classifier's ability to induce its model. To test this we measured the number of fluctuations in activity in each community - i.e. going from *decrease* at one time step to *increase* at the next or vice-versa - and correlated this with the accuracy measures (precision, recall and f-measure) using the Pearson correlation coefficient (r). We found precision and the fluctuation count to be negatively correlated ($r = -0.514$ where $p < 0.001$) indicating that as the number of fluctuations increases precision reduces and that recall and fluctuation count were also negatively correlated ($r = -0.589$ where $p < 0.01$) indicating a similar association between fluctuation increase and performance reduction.

Given that the logistic regression model yielded the best performance (i.e. outperforming the J48 classifier in terms of the F_1 level) we then assessed the model's performance when detecting activity decreases - given that these are of concern to community operators - by plotting the Receiver Operator Characteristic (ROC) curve for each community. Figure 13 presents the ROC curves, showing the tradeoff between the True Positive Rate (TPR) - i.e. recall - and the False Positive Rate (FPR) for each community's logistic regression model. It demonstrates that using the role composition we can accurately predict a decrease in community activity for 23 of the 25 communities analysed - i.e. by surpassing the random predictor given by the grey line running from the bottom-left corner to the top-right. The two communities that the predictions were worse than the random classifier were **319 (Best Practice and Bench Marking)** and **210 (Analytics)**. In each case we find that the kappa statistic (κ) of the class agreement is negative - -0.075 and -0.70 for 319 and 210 respectively - suggesting that the role composition in these communities provides little information for the class predictions.

6.4. Results: Post/User Count Regression

For our third experiment we performed two regression analysis tasks. The first analysis honed in on the differences between communities by exploring the question *Do distinct communities exhibit disparate patterns in how role compositions affect community activity?* We were interested in assessing how communities differ from one another in the relationship between behaviour in the communities and activity, and explored this

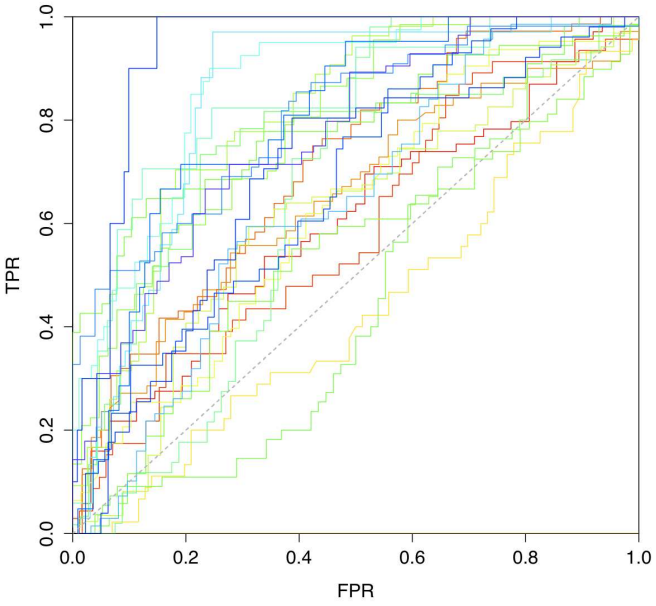

Figure 13: ROC plot for activity decrease detection (from the previous time step) when using logistic regression trained on a community's role composition. The random predictor is given by the grey line running from the lower-left corner to the top-right corner.

correlation by inducing linear regression models for each community that predicts the community's post count. The *dependent variable* was set as the post count from the *feature window* and each role was used as an *independent variable* with the composition percentage used as the value of the variable. For the second analysis we explored the relationship between compositions across the entire platform and the number of active users present within the communities. The reasoning behind this analysis was to explore the relation between community size and compositions, thereby identifying any roles that were more prevalent within larger communities. We performed regression analysis by setting the user count in the *feature window* as the *dependent variable* and each role as an *independent variable* with the composition percentage as the variable's value.

6.4.1. Post Count Regression

Figure 14 shows the PCA plot for each community using the regression model's coefficients as the *composition motif*. Unlike in the previous PCA plot, in Figure 10 for the average role compositions, in this case the communities are not as greatly dispersed. Instead we find that forums **50 (ABAP General Discussion)** and **419 (SAP Business One System Administration)** are isolated, whereas before the former community was clustered near **256 (Governance, Risk and Compliance)**. This indicates that although the average compositions may be similar between forums, what is correlated with activity is in fact different. Figure 14 also demonstrates that there is a large central cluster where the coefficients from the regression models are all similar. Within this tight cluster we find forum **470 (Manufacturing Execution)** which was found to have a distinct average composition in our first experiment.

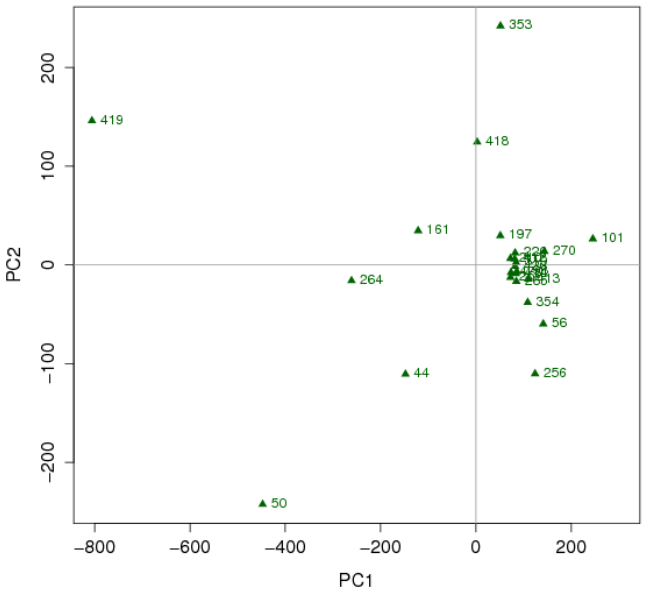

Figure 14: Principal Component Analysis (PCA) plot of each community where linear regression coefficients are used as the *composition motifs*.

To provide a greater insight into how the communities' compositions differ we assessed the linear regression models' coefficients of seven communities selected from Figure 14, choosing those communities that are dispersed or representative of clustering in the plot:

- 50 (ABAP General Discussion)
- 264 (SAP Business One Core)
- 353 (SAP Business One Reporting & Printing)
- 419 (SAP Business One System Administration)
- 44 (Process Integration)
- 56 (SAP Business One SDK)
- 270 (Financial Performance Management)

Table 5 presents each of the seven communities' regression model coefficients and their significance levels assessed using the t-test. Commonalities exist across the communities in terms of the importance of certain roles. For instance for **Focussed Expert Initiator** we find that for 50, 264 and 44 an increase in this role is associated with increased activity, while for 270 a decrease in this role is correlated with an increase in activity. For **Focussed Expert Participant** we find that for all the communities an increase in this role is correlated with an increase in activity.[19]

[19]We only consider features that are significant within the model.

Table 5: Model coefficients and the adjusted coefficient of determination (R^2) values for seven SAP communities' linear regression models. The model regressed the post count on the nine roles and their composition values.

Role	**50**	**264**	**353**	**419**	**44**	**56**	**270**
Focussed Expert Participant	15.292	191.938***	159.255**	529.200***	2.200	-21.268	-2.804
Focussed Novice	1.462	-64.370***	58.423**	11.235	2.484	-68.126***	-11.454***
Mixed Novice	2.437	6.966	113.883***	46.983*	2.161	-63.295***	-19.755***
Distributed Expert	4.678	-0.929	-15.780	-4.313	4.180	-38.775***	2.343
Focussed Expert Initiator	588.277***	290.651***	-55.581	721.285	257.895***	-44.844**	-64.787***
Distributed Novice	4.537	-40.871*	21.586.	10.995	1.058	-57.014***	-13.107***
Knowledgeable Member	1.119	-84.804*	113.275***	29.726	1.835	-70.283***	-6.664.
Mixed Expert	12.505.	10.997	5.852	81.802**	5.374	-51.882***	-15.099.
Knowledgeable Sink	-1.831**	-47.463.	146.377***	56.592**	6.288*	-59.661***	9.244*
Adjusted R^2	0.974	0.768	0.604	0.435	0.949	0.916	0.927

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1

The models' coefficients also indicate, in general, that an increase in novice users is associated with a decrease in activity. For instance for **Mixed Novice** - i.e. a user who is a non-expert and whose topical focus has a medium distribution (neither distributed nor focussed) - forums 56 and 270 have negative coefficients for this role, while for **Distributed Novice** forums 264 and 270 are found to have negative coefficients.

6.4.2. User Count Regression

The post count regression analysis demonstrated the idiosyncratic patterns that appear in each community and the unique dependencies between role compositions and community activity. As forums differ in their size and scale, one question that was provoked from this analysis was whether a relationship existed between the size of a forum and the composition that it exhibits. One would presume that forums with many users require mediating users who participate by both initiating threads and joining in existing discussions. To assess this we performed a second regression analysis task, this time by assessing all communities in the SCN in a single regression model that regressed the user count on the role composition.

Table 6: Coefficients from the linear regression model where the user count is predicted using the role composition of the SAP communities, and the Pearson correlation coefficient between the user count and the role compositions in the dataset.

Role	Regression Coefficient	*r*
Focussed Expert Participant	26.528 ***	0.254***
Focussed Novice	0.553 ***	0.133***
Mixed Novice	0.050	-0.076***
Distributed Expert	0.225	0.002
Focussed Expert Initiator	-5.313	-0.063***
Distributed Novice	-0.164	0.095**
Knowledgeable Member	-2.899**	-0.046*
Mixed Expert	-0.897	0.011
Focussed Knowledgeable Sink	8.756***	0.267***
Adjusted R^2	0.114	-

Signif. codes: p-value < 0.001 *** 0.01 ** 0.05 * 0.1 . 1

Table 6 shows the coefficients that our regression model produced and the coefficient of determination of the model with the data. The results indicate a weak fit (i.e. $R^2 = 0.114$) to the data, suggesting that this regression model cannot describe the relation between forum size and composition in a meaningful way. That said, in the model we do find statistically significant features. For instance, we find that an increase in **Focussed Expert Participants**, **Focussed Novices** and **Knowledgeable Sinks** is associated with an increase in the user count. This in line with our earlier presumption that forums with larger numbers of users require roles that both initiate and participate within existing discussions.

To assess the dependencies between community size and individual roles, we also measured the Pearson correlation coefficient (r) between the user count and the roles in the dataset. We found that the coefficients are relatively low and do not explain any strong relations between an increase in the user count and the roles. The highest is **Knowledgeable Sink** which suggests that this role is more prevalent in forums with larger numbers of users, given that this role is associated with heightened *popularity*.

7. Discussion and Future Work

Our three-stage approach for the role composition analysis of online communities functions by a) *modelling* user behaviour and roles, b) *identifying* roles on a given community platform and c) *analysing* community health using role compositions. We now discuss the issues and findings from each stage.

7.1. Modelling

The presented behaviour ontology extends SIOC and is capable of representing the contextual notion of behaviour where the same user may exhibit differing behaviour within different time periods or localities. Our method for labelling users with their community roles employs semantic rules, in particular SPIN functions, that are constructed from a given *Skeleton Rule Base* - where this rule base is constructed depending on the platform in question, for example by including a set of roles to match discussion-based roles for a discussion message board. By using dynamic binning we were able to account for fluctuations in community behaviour and, more importantly, enable our rule-based approach to be applied over iterative time steps. Existing statistical approaches to composition analysis [4, 5, 8, 6] require cluster centroids to be mapped to one another between time steps, thereby preserving the role labels for the clusters. Without such mapping, repeated clustering must be performed and an analyst involved within the process in order to assign the role labels to clusters.

Our future work will involve the exploration of *role life cycles* to model the movement that users exhibit between roles within communities. In doing so we can devise a probability-based framework in which the transitional likelihood of a user moving from one role to another can be derived. This would in turn support community managers in tracking the role development of individual users and identifying which users are likely to turn into community leaders or experts and, more importantly, which are likely to churn. This work is eased through the use of semantic web technologies given that we now have examples of role life cycles according through our behaviour ontology.

A second avenue of future work will be to extend our behaviour ontology for various community types. The current version of the ontology forms a *Core* specification for contextual behaviour at a generic level. Our future work will provide platform-specific extensions of this ontology for roles that we have identified for a given platform and machine-readable descriptions of feature derivation techniques for each of the aforementioned behaviour dimensions. SCN will provide the starting point for this.

7.2. Role Identification

The described method for rule tuning uses statistical clustering methods to achieve the optimum partitioning of users into behavioural clusters before aligning those clusters with role labels through a maximum-entropy decision tree. This decision-tree method chooses the paths of shortest depth through the tree and from this generates the role label to use for each cluster. In our previous work [20] we assessed the role compositions of three distinct community forums from the Boards.ie message board platform and yielded an unclassified user rate of 29%, however using our maximum-entropy decision tree we now yield a reduced unclassified user rate of 7%. The improvement in reducing the number of unclassified users is due to the nature through which our maximum-entropy decision tree method yields the role labels, as it selects the dimension that generates the *purest* split at each decision node. Our previous work in a similar vein to existing work [8, 6] however relied solely on the manual projection of role labels, from the literature, to clusters without this intermediary *role identification* step that grounds the roles to the platform. We anticipate that this approach for tuning the roles to a given community will be of great use to analysts who wish to derive the role composition for their community platform. Our future work will involve applying our role identification method over Boards.ie, Twitter and other community platforms to derive the role labels that are relevant in those contexts.

7.3. Analysis

By exploring the three research questions defined within the introduction of this paper we found the analysed communities to exhibit both commonalities and idiosyncrasies. For instance when exploring *What roles are dominant in disparate communities?* we found novice users to be common across all the communities and that the discriminating factor between the forums was the *focus dispersion* of such users being either *low, mid* or *high*. We also found communities to differ in terms of the experts who participated in a similar manner to the separation of novice users based on focus dispersion. This suggests that failing communities which share common topics could have certain expert users brought in, particular if their past health was related with the inclusion of *Distributed Experts*.

Through addressing the question *Do distinct communities exhibit disparate patterns in how role composition affects community activity?* our analyses also identified differences in the association between the proportion of novice users and activity within communities, where in certain forums an increase in novice users was linked to an increase in activity while being the opposite in others. In our previous work [20], when analysing three different community forums from Boards.ie, we found similar idiosyncratic properties where the role *supporter* - designating a user who joins discussions but who does not initiate them - was negatively associated with activity in one forum while there was no correlation, neither positive nor negative, for the two remaining forums.

Such insights have provoked two pertinent questions, firstly *is the role composition of a community simply a reflection of its type?* And *are the results simply due to the type of people that join the community?* If we can understand this distinction then we can provide a better insight into whether the community is healthy or not - i.e. tailoring a health metric based on the community type or assessing the value of individual users to the community. Future work will explore these two questions, seeking the distinction between the migration of types of users and the type of the community.

Using the roles and their composition percentages we were able to detect either an increase or a decrease in community activity through a binary classification task - addressing the research question *How does a change in its role composition affect the community?* We found that this approach was able to outperform a random selection baseline for 23 out of the 25 analysed communities. Measuring the number of activity change fluctuations within each community gave an indication as to how often the post count varied from week-to-week. Using this information we found a negative correlation between the fluctuation count in a community and the accuracy of the logistic regression model in terms of both precision and recall. This indicates that for communities in which activity changes often that the role composition of the community may not carry sufficient information to facilitate the detection of such changes.

8. Conclusions

The widespread uptake, usage and provision of online communities by companies and organisations means that there is a vested interest in such communities remaining healthy and active. In communities users interact with one another around a shared topic or interest and exhibit behaviour that can be used to label them with their roles in the community. By deriving the role composition of a community - i.e. the percentage distribution of different roles - the composition can be associated with

signifiers of health, such as activity, and used to identify what worked for the community and what did not.

In this paper we have presented a three-stage approach to facilitate the process of community health analysis through: a) the *modelling* of user behaviour, b) the *identification* of roles that are relevant to a given platform, and c) the *analysis* of a community's health based on its role composition. We presented an ontology to model user behaviour that captures the notion of disparate behaviour within differing contexts - i.e. time and location - and a dynamic approach to infer the role of a user based on his/her exhibited behaviour with semantic rules. We described a method to tune roles to a specific community using statistical clustering and discretisation, and also introduced a novel means to derive role labels for clusters using a maximum-entropy decision tree. Finally, we demonstrated the utility of deriving the role composition for a community by: a) identifying differences between communities, b) accurately detecting activity changes, and c) accurately predicting community activity, all using a community's role composition derived from behaviour dimensions and semantic rules.

9. Acknowledgements

The work of the authors was supported by the EU-FP7 project Robust (grant no. 257859). We would also like to thank SAP for the provision of the dataset for our analyses.

References

[1] J. Preece, Online Communities - Designing Usability, Supporting Sociability, John Wiley & Sons, Ltd, 2000.

[2] S. A. Golder, J. Donath, Social roles in electronic communities, in: in Association of Internet Researchers (AoIR) 5.0, 2004, pp. 19–22.

[3] J. Hautz, K. Hutter, J. Fuller, K. Matzler, M. Rieger, How to establish an online innovation community? the role of users and their innovative content, in: System Sciences (HICSS), 2010 43rd Hawaii International Conference on, 2010, pp. 1 –11. doi:10.1109/HICSS.2010.221.

[4] R. Nolker, L. Zhou, Social computing and weighting to identify member roles in online communities, in: Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, 2005, pp. 87 – 93. doi:10.1109/WI.2005.134.

[5] T. Zhu, B. Wu, B. Wang, Social influence and role analysis based on community structure in social network, in: Proceedings of the 5th International Conference on Advanced Data Mining and Applications, ADMA '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 788–795.

[6] T. Zhu, B. Wang, B. Wu, C. Zhu, Role defining using behavior-based clustering in telecommunication network, Expert Syst. Appl. 38 (2011) 3902–3908. doi:http://dx.doi.org/10.1016/j.eswa.2010.09.051.
URL http://dx.doi.org/10.1016/j.eswa.2010.09.051

[7] J.-W. Strijbos, M. F. D. Laat, Developing the role concept for computer-supported collaborative learning: An explorative synthesis, Computers in Human Behavior 26 (4) (2010) 495 – 505, emerging and Scripted Roles in Computer-supported Collaborative Learning. doi:DOI: 10.1016/j.chb.2009.08.014.
URL http://www.sciencedirect.com/science/article/pii/S074756320900137X

[8] J. Chan, C. Hayes, E. M. Daly, Decomposing discussion forums and boards using user roles, in: ICWSM, 2010.

[9] D. Fisher, M. Smith, H. Welser, You are who you talk to: Detecting roles in usenet newsgroups, in: System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on, Vol. 3, 2006, p. 59b. doi:10.1109/HICSS.2006.536.

[10] M. Maia, J. Almeida, V. Almeida, Identifying user behavior in online social networks, in: Proceedings of the 1st Workshop on Social Network Systems, SocialNets '08, ACM, New York, NY, USA, 2008, pp. 1–6. doi:http://doi.acm.org/10.1145/1435497.1435498.
URL http://doi.acm.org/10.1145/1435497.1435498

[11] E. Gleave, H. T. Welser, T. M. Lento, M. A. Smith, A conceptual and operational definition of 'social role' in online community, in: HICSS, 2009, pp. 1–11.

[12] J. G. Breslin, A. Harth, U. Bojars, S. Decker, Towards semantically-interlinked online communities, in: Proc. 2nd European Semantic Web Conf. (ESWC), Springer, 2005, pp. 500–514. doi:10.1007/11431053_34.

[13] A. Ankolekar, K. Sycara, J. Herbsleb, R. Kraut, C. Welty, Supporting online problem-solving communities with the semantic web, in: Proceedings of the 15th international conference on World Wide Web, WWW '06, ACM, New York, NY, USA, 2006, pp. 575–584. doi:http://doi.acm.org/10.1145/1135777.1135862.
URL http://doi.acm.org/10.1145/1135777.1135862

[14] D. Brickley, L. Miller, FOAF Vocabulary Specification 0.97, Namespace document, W3C (January 2010).
URL http://xmlns.com/foaf/spec/20100101.html

[15] P. Mika, Ontologies are us: A unified model of social networks and semantics, Web Semant. 5 (2007) 5–15. doi:10.1016/j.websem.2006.11.002.
URL http://dl.acm.org/citation.cfm?id=1229184.1229195

[16] R. Hoekstra, Representing social reality in OWL 2, in: 7th International Workshop on OWL: Experiences and Directions (OWLED 2010), 2010.

[17] M. Rowe, S. Angeletou, H. Alani, Predicting discussions on the social semantic web, in: ESWC (2), 2011, pp. 405–420.

[18] T. K. Moon, The expectation-maximization algorithm, IEEE Signal Processing Magazine 13 (6) (1996) 47–60. doi:10.1109/79.543975.
URL http://dx.doi.org/10.1109/79.543975

[19] S. P. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory 28 (1982) 129–137.

[20] S. Angeletou, M. Rowe, H. Alani, Modelling and analysis of user behaviour in online communities, in: L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist (Eds.), The Semantic Web – ISWC 2011, Vol. 7031 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 35–50.